Information Extraction from EHR

USING NLP TECHNIQUES - AN EXPLORATION

NAYAN CHAUDHARY

Contents

- Problem Statement
- Possible approaches
- Introduction to tool
- Some explorations
- Challenges

Disclaimers

- Have to guard many specifics about the data – its proprietary
- Extensively domain-specific knowledge
- This isn't a dummy dataset, so no "known truth/baseline" results
- No annotations whatsoever
- Time/Resource/business constraint
- Results still exploratory

Data characteristics

- Transcribed physician notes, free text with redactions
- Some longitudinal component in the data
- Data from a very specific patient subset :
  - All patients are cancer patients, thus oncology data
  - All have been treated with a very specific class of onco-drugs
- Data size :
  - ~ 11,500 records
  - ~ 5,500 patients

Example

CHIEF COMPLAINT:<LGMX_BR/>

The patient was requested by **NAME[5P8E606PG]. Patient known to you **with history of melanoma**, now with brain lesion.<LGMX_BR/>

<LGMX_BR/>

HISTORY OF PRESENT ILLNESS:<LGMX_BR/>

ID[1ZBZCPR21]. G77XDX is a 62-year-old very pleasant gentleman with history of alcoholic cirrhosis of the liver, **recently was diagnosed with metastatic melanoma to the lung. He was started on xxxxxxxxx in 04/2016. He was brought in by his wife to the ER complaining of altered mental status, slurred speech for the past 2 days. He has not being eating for the past 2 days. Denies any fevers or night sweats. In the ER, he underwent CT of the head, which evidently showed large **DATE[01/23]cm hyperdense lesion in the left parietal lobe with extensive surrounding edema. There may be a small lesion in the right posterior parietal lobe with adjacent edema. There is a small mass effect on the left lateral ventricle. He received Decadron 10 mg IV in the ER. Today, we are consulted for further management. He is more alert now. His wife is at the bedside. She does mention that since he got a steroid, he was able to talk and also able to eat food. Denies any chest pain or difficulty breathing. Denies any abdominal pain. Bowel movements have been regular. Denies any hematochezia or melena. Denies any dysuria or hematuria. Denies any headaches or lightheadedness.<LGMX_BR/>

<LGMX_BR/>

PAST MEDICAL HISTORY:<LGMX_BR/>

Diabetes, anemia, hypertension, history of alcoholic cirrhosis of the liver, thrombocytopenia secondary to hypersplenism, recurrent right plantar foot ulcer with infection, chronic anemia, history of osteomyelitis.<LGMX_BR/>

Problem Statement

- Key questions (eventual) :
  - Identify 'causal' {Drug, Adverse Events} pairs – a relationship extraction task
  - If Adverse Event, subsequent course of action like changes in therapy, drugs, treatment discontinuation etc.
- Other information extraction tasks:
  - Primary cancer being treated for
  - Line of Therapy
  - Stage of Disease
  - Treatment discontinuation, reasons
  - ……..

Adverse events ?

- Side effects to using cancer drugs

List of some immune-related Adverse Events	
Vitiligo	Thyroiditis
Uveitis	Hypothyroidism
Myocarditis	Hyperthyroidism
Pancreatitis	Pneumonitis
Autoimmune Diabetes	Thrombocytopenia
Colitis	Anemia
Enteritis	Hepatitis
Encephalitis	Adrenal insufficiency
Aseptic meningitis	Nephritis
Hypophysitis	Vasculitis
Neuropathy	Arthralgia

What's causal ?

- *"I recommended xxxxxxx and informed the patient of possible side effects, such as colitis, but he declined"*
  - This is just a 'mention' of colitis, in the context of the drug 'xxxxxx' but not an actual patient event
- *"Following xxxxxxx, the patient experienced severe colitis"*
  - This is an actual event of 'colitis', that happens in the context of the drug 'xxxxxx'
  - This isn't statistically causal yet, but 'causal' for our purposes

Solution approaches

Code-based	Linguamatics I2E

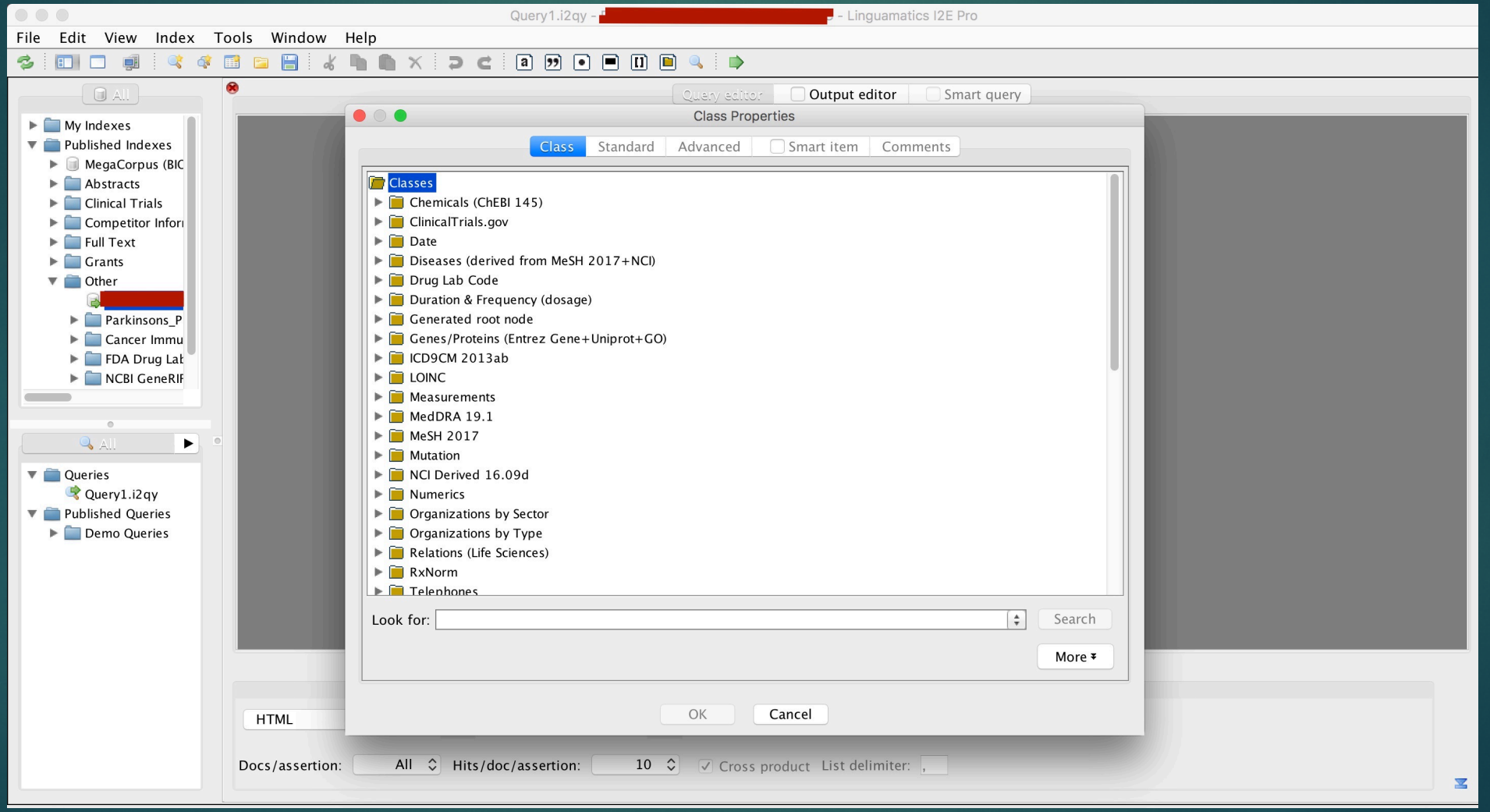
• regular algorithmic approaches supervised/semi-supervised
  • No labeled data, no annotations
  • Medical/Biological context terms/dictionaries
	• Proprietary licensed software with in-built NLP algorithms
	  • Suitable for bootstrapping-type approach to relationship extraction
	  • Extensive medical/biological dictionaries Eg. ICD codes

Had to learn the tool*

Linguamatics I2E

- A text-mining tool for unstructured text
- Has support for writing very complex queries (classes, verbs, alternatives etc.)
- Can build domain knowledge in by writing macros
- Good for visualization
- Has APIs as well to access using many programming languages but haven't explored that

Linguamatics I2E



Primary Cancer Site - exploration

- Primary cancer site is tricky in cases:
  - Patients having history of multiple cancers
  - Metastasis
  - Overlapping/ambiguous cancers
- Focus on:
  - Lung, Skin, Kidney, Colorectal, Liver, Gastric, Bladder
- What ‘class’ to use ?

Evaluation – Primary Cancer Site

Very generic queries	Very specific queries
• return a hit/answer for all documents • many cancer sites might be returned	• Only a few documents might have that pattern • no cancer sites captured

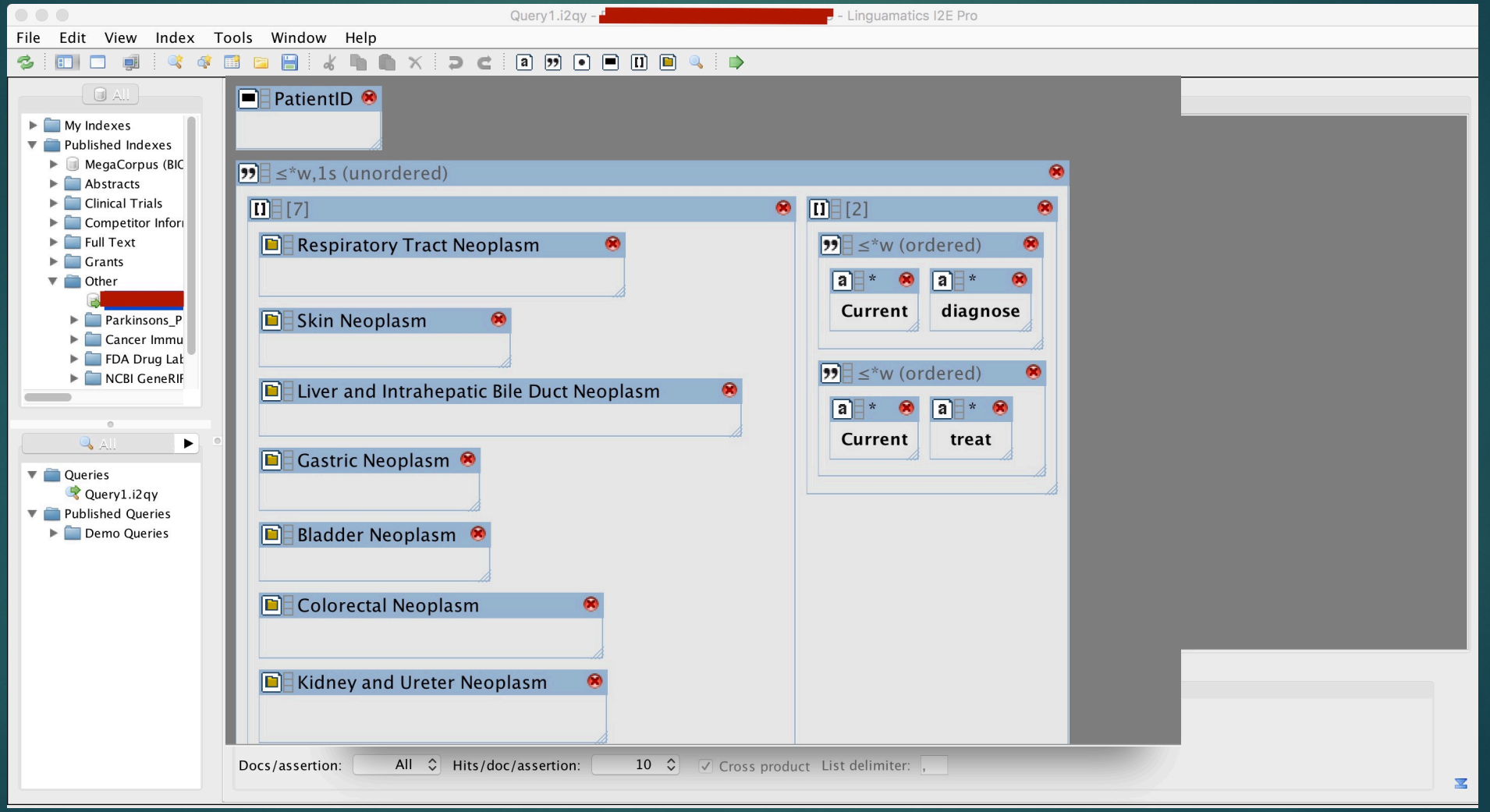
- Hence following metrics:
  - **% Answered** – For how many Patient IDs query has an answer
  - **Recall** – Among answered, if true value exists in the list/value returned
  - **Precision** – Among answered, if true value and value returned is same

Result Examples – Primary Cancer Site

Cancer Mentions	Preceding verbs - Current Diagnose, current treat
• % Answered – 100% • Recall - 0.90 • Precision – 0.28	• % Answered – 10% • Recall - 1.0 • Precision – 0.80

- Test Set – 50 patients
- Create/choose classes specific to cancer sites we are interested in
- Sentence pattern where verbs alternatives precede the cancer mention

Linguamatics I2E query example



Challenges/Open questions

- Annotations for evaluation:
  - How to choose documents ?
  - How to guarantee coverage ?
  - Focus on precision/recall ?
  - Do it class specific ?
- How tractable is the task in this unsupervised setting ?